Lang-8を用いた日本語学習者向けの誤用検索システムの構築

新井 美桜　小平 知範　小町 守
首都大学東京
cherry.m@mac.com, kodaira-tomonori@ed.tmu.ac.jp, komachi@tmu.ac.jp

1　はじめに

近年、日本語を母語としない人々の日本語学習の需要が高まっている。文化庁の調査では[1]平成2年度では60,601人であった国内の日本語学習者は平成27年度では191,753人にまで増加している。また、国際交流基金によれば[2]、平成27年度の海外での日本語学習者は3,651,715人であった。そのため、日本語学習を支援するための研究が注目され始めている。

学習者の作文を支援するための方法として、用例検索システムが挙げられる。用例検索システムとは多数の文を収録したコーパスから探したい語句を含む文を検索し、どのようにその語句が使われているか確認することによって学習者の作文を支援するシステムである。

学習者は、作文を執筆する際、用例検索システムで使用したい語句がどのように使われているかを確認する。学習者の検索したい語句が正しければ、通常の用例検索システムでも検索できる。しかし、学習者の想像している語句が間違っている場合、用例検索をすることができない。また、誤用の検索ができたとしても、それに対する添削がなければ、それがどのように間違っていて、どのように直せば良いのかわからない。

しかしながら、既存の日本語学習者の作文用例検索システム [1][2] では、正用例のみまたは誤用例のみの検索しかできない。また、用例の総数が少なく、学習者が検索したい語句を入力しても、十分な情報を取得できない場合がある。また、既存の日本語用例検索システムは、主に日本語教育者向けに作られたものが多く、日本語の予備知識がない人は活用できない。

本研究では、誤用例の検索の有用性に着目し、日本語学習者を対象として、相互添削型SNS、**Lang-8**[3]**を用いた日本語学習のための大規模誤用検索システム**を構築した。Lang-8のデータセットは誤用文に日本語母語話者の書いた添削文が付与されており、このデータセットを誤用例検索に利用することで、学習者の文がどのように間違っていてどのように直せばよいのかの判断の参考にできる。また、係り受け構造を利用し、共起表現を表示することで、どんな表現がよく使われるのか、そしてどのように使用されるのかを、よりわかりやすく学習者に提示できるようにした。また、すべての用例に誤用文が付与されているため、学習者がどの語句をどのような形で間違えやすいのかを誤用例検索で知ることができ、言語教育に利用することができる。

この検索システムはWebサイト[4]にて公開している。

2　関連研究

今日、様々な日本語学習者向けの用例検索システムが開発されている。しかし、未だ学習者が実用的に利用することは難しい部分が多く存在する。例えば、既存の日本語学習者の作文用例検索システムには以下のようなものがある。

東京外国語大学のコーパスに基づく言語学教育研究拠点はEラーニングを活用した『日本語学習者言語コーパス』[5]を構築した。『日本語学習者言語コーパス』は検索したい学習者の母語や年齢、性別などを選択するとその条件にあった文を検索ワードとともに検索でき、検索したキーワードに一致した文がKWIC (keyword in context) 形式で表示される。日本語教育者はこの検索システムを使うことで、学習者の誤りの特徴を得ることができる。この検索システムは学習者というよりは教育者を対象にしたもので、学習者が扱いやすいような用例の表示の仕方はしておらず、また、誤用検索や対応する添削文があるわけではないので、学習者の検索したキーワードが誤っていた場合、検索

[1] http://kodomo.bunka.go.jp/koho_hodo_oshirase/hodohappyo/pdf/2016072801_besshi01.pdf
[2] https://www.jpf.go.jp/j/about/press/2016/dl/2016-057-2.pdf
[3] http://lang-8.com
[4] http://cl.sd.tmu.ac.jp/nihongo
[5] http://cblle.tufs.ac.jp/llc/ja/search.php?menulang=ja

することができない。また海外の大学の日本語学科の学生が書いた作文を収集し、『日本語誤用コーパス』[6]を構築し、誤用検索機能をつけて公開したが、総文数が 654 文と非常に少なく、検索できない語が多い。

李ら [1] は日本語学習者の発話データを文字化した言語資料である『KY コーパス』に言語情報を付与した『タグ付き KY コーパス』に用例検索機能を追加し、公開した。『タグ付き KY コーパス』は検索したい語を入力するとコーパスから学習者の正用例、誤用例を検索できる。こちらも誤用例は非常に数が少なく、頻繁に使用するであろう語句にも誤用例が存在しない場合が多くある。また誤用例に対応する正用例がないので、学習者が作文に一致する誤用箇所を見つけても、訂正の仕方がわからず、学習者が参考にすることは難しい。

Hinoki プロジェクトの日本語共起語検索システム『なつめ』[2][7]は名詞、動詞、形容詞を選択し、単語を入力して検索すると、その語と共起しやすい助詞と語（名詞で検索した場合は動詞や形容詞、動詞や形容詞で検索した場合は名詞）を頻出順に表示する。学習者は、どの語が共起しやすいのか確認することで、自分の作文の参考にすることができる。また、共起語を正しい助詞とともに表示するので、日本語学習者にとって最も困難である助詞の使用方法を確認することができる。しかし、このシステムには具体的な用例を表示する機能がないので、頻繁に使用される語の組み合わせがわかっても実際に利用するのは難しい。

表 1 に、先に述べた関連研究の検索システムが利用できる機能を示す。正用は正用検索の有無、誤用は誤用検索の有無、添削は誤用文に対する添削の有無、共起は共起表現の提示があるかどうかを表す。それぞれ、利用できる場合は○、できない場合は×で表している。

本研究では、Lang-8 データセットを用いることで、既存の日本語学習者の作文の用例検索と比較して超大規模な正用例検索、添削文付きの誤用例検索を可能にした。また、『なつめ』と同様に共起表現を用いた検索機能も追加した。

3 Lang-8 データセット

Lang-8 は、言語学習者向けの相互添削型 SNS である。学習者が学習している言語で日記を書くと、その言語の母語話者が添削をしてくれる。反対に、自分の母語の学習をしている人の日記の添削もできる。修正

[6]http://cblle.tufs.ac.jp/llc/ja_wrong/index.php?m=default
[7]https://hinoki-project.org/natsume/

システム名	正用	誤用	添削	共起
日本語学習者言語コーパス	○	×	×	×
日本語誤用コーパス	×	○	○	×
タグ付き KY コーパス	○	○	×	×
なつめ	×	×	×	○
本検索システム	○	○	○	○

表 1: 関連研究との比較

箇所には添削者によって任意で定められた、[sline]、[f-red]、[f-blue] などのタグ付けがされている。

本研究では水本ら [3] によって作成された学習者コーパス、Lang-8 Learner Corpora を使用した。Lang-8 Learner Corpora は Lang-8 利用者の 2007 年から 2011 年までの作文データが JSON 形式で収録されており、学習者の作文とその添削、エッセイ ID、ユーザ ID、学習言語タグ、母語タグからなる。学習言語のタグが Japanese のエッセイ数は 185,991 である。

本研究では、Lang-8 Learner Corpora の日本語学習者が書いた添削前の誤用例と日本語母語話者が添削した正用例のペアのセット 140 万文対を抽出した。

4 誤用検索システム

本節では、本検索システムを構築する際に行った前処理と実装、および本検索システムのアルゴリズムおよびインターフェースについて説明する。

4.1 前処理

Lang-8 Learner Corpora のデータは 1 つの学習者の文に 1 文から複数の添削文が付与されている。それぞれを 1 つの文対にするため、1 つの学習者文と 1 つの添削文をセットにした。そのため、学習者の文が同じであっても、添削が異なる場合は、別の文対とした。また、複数の添削文の中で添削が同じものは 1 つにまとめた。余分な情報が入っている文を取り除くため、学習者の書いた文の文長が 100 文字以上のものと学習者の文と添削文のレーベンシュタイン距離が 7 より大きいものは除去した。

使用する学習者と添削文の文対は単語で分かち書きをした。形態素解析器には MeCab (ver 0.996)[8]を使用した。学習者は誤った文を書く可能性が高く、解析ミスが起こりやすいため、辞書は短単位で分割できる UniDic (ver 2.2.0) を使用した。CaboCha (ver 0.69)[9]

[8]https://github.com/taku910/mecab
[9]https://github.com/taku910/cabocha

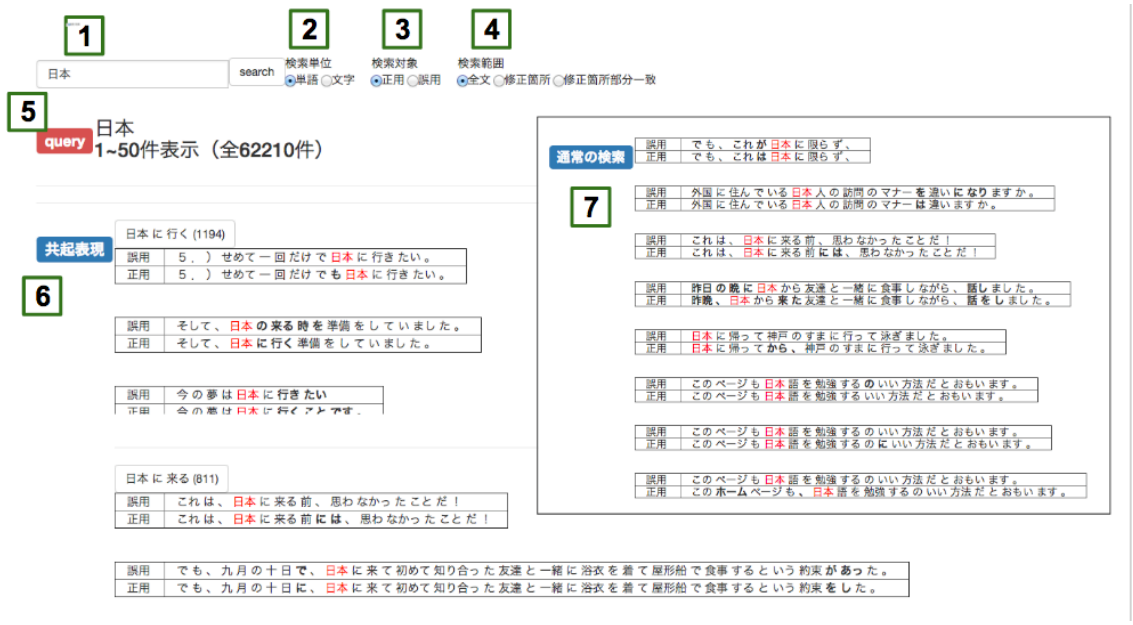

図 1: 誤用検索システム外観

を使用し、係り受け関係にある文節対から係り元にある名詞（主辞）と助詞、係り先の文節の動詞（主辞）の 3 つ組を添削後の文から抽出し、頻度を数えた。これらを共起表現として扱う。

4.2 用例検索アルゴリズム

用例のデータベースには 3 節で構築したデータを用いる。入力クエリは単語または句（連続する単語列）を想定している。単語単位の検索のときはクエリに対して形態素解析を行うが、解析ミスが起きる可能性があるためユーザの任意で文字単位に変更が可能である。検索対象は誤用または正例のどちらか一方がユーザによって決められる。検索のアルゴリズムとしては、対象とする文中の文字列または単語列とクエリが完全一致するもののみを抽出する。完全一致の対象は文の全文、誤用箇所に部分一致、誤用箇所に完全一致の 3 種類いずれかである。

4.3 検索インターフェース

本検索システムの外観である図 1 の説明を以下に示す。

1. 検索窓: 検索したい語句を入力する。
2. 検索単位: 分かち書きの単位を、単語か文字で選択する。デフォルトでは単語である。

番号	問題
1	あなたの国の行事について
2	人生で最も心に残る冒険
3	日本の好きなところ
4	最も印象に残った映画／本について
5	あなたの国の料理について
6	世界共通語としての英語の是非について

表 2: 評価用問題の一部

システム名	評価者 A	評価者 B
KY コーパス	2	5
本検索システム	**5**	**8**

表 3: 客観評価（評価者の日本語作文の点数）

3. 検索対象: 検索する対象を正用例か誤用例で選択する。デフォルトでは正用である。
4. 検索範囲: 検索する文の範囲を、全文、修正箇所、修正箇所部分一致の中から選択する。全文を選択すると文の全体を、修正箇所を選択すると修正箇所のみでクエリに一致する文を、修正箇所部分一致を選択すると修正箇所に部分的に一致する文を検索する。デフォルトでは全文である。
5. クエリと件数表示: 検索したクエリと検索件数が表示される。検索単位を単語にした場合、クエリが単語区切りで表示される。1 ページにつき最大 50 件表示される。

質問	回答者	回答
KY コーパスと本システムの違い	A	KY の方が複雑で、使い方はよくわかりませんでした。本検索システムは優しいと思います。
	B	前者の検索システムは検索できるものは限られるが、後者の方が範囲がもっと広いです。しかも、後者の例文も多くて使いやすいです。
今後の改善点	A	複数の単語で検索できること。
	B	全文の文法的な間違いがあるかどうかを検索する機能があったらいいと思います。

表 4: 主観評価（評価者によるアンケート回答結果）

6. 共起表現: 名詞もしくは動詞で検索する際、クエリに含まれる名詞または動詞と共起する名詞、助詞、動詞の組み合わせが頻度順に最大 10 件表示される。この組み合わせにはそれぞれ下方に、その表現が使われている用例が表示されており、スクロールすると最大 10 件の用例を見ることができる。

7. 通常の検索結果: 選択した条件に一致した検索結果が共起表現の表示の下に表示される。上に学習者の書いた誤用例、下に誤用例が添削された正用例がペアになって表示される。クエリの部分は赤く、修正箇所は太字で表示される。

5 誤用検索システムの評価実験

評価は、誤用例と正用例の文対表示の実用性を確かめるため、『タグ付き KY コーパス』の用例検索システムと本システムを比較した。日本語作文の問題を 10 問用意し、日本語を母語としない評価者 2 人に回答してもらった。評価者 A には、奇数番号の問題を『KY コーパス』を使って、偶数番号の問題を本システムを使って解いてもらった。評価者 B には偶数番号の問題を『KY コーパス』を使って、奇数番号の問題を本システムを使って解いてもらった。評価基準を統一するため、1 問につき 3 文の制限を付けた。結果を以下の 2 つの評価方法で評価した。

客観評価　客観評価では、評価者に書いてもらった文章に減点法で点数をつけた。持ち点 1 人 10 点で、1 つの文法誤りにつき 1 点引いていき、各システムごとに 5 問ずつ合計し、最終的な点数を計算した。その結果を表 3 に示す。両者ともに、本検索システムの方が点数が高く、日本語作文支援に有効であることが示された。

KY コーパスを用いた作文では、単純な文法誤りや語法の誤りが多く、本検索システムを用いた作文では、助詞の誤りや語の欠落が多かった。KY コーパスは誤用文の添削がなく、本検索システムは品詞の表示がないため、このような誤りが多くなったと考えられる。

主観評価　主観評価では、評価者個人に、2 つの検索システムがどのように異なり、また、どちらの検索システムが使いやすかったのかをアンケートした。その結果を表 4 に示す。主観評価においても本検索システムが有用であることが確認できた。

6 おわりに

本研究では、作文執筆の際に学習者が直面する困難に着目して、Lang-8 の日本語学習者コーパスを利用し、誤用検索システムの構築を行った。検索システムでは、添削文付きの誤用例文を表示した。さらに、係り受けを利用した高頻度の共起表現を含むような用例の提示を行った。検索システムの評価実験では、学習者の日本語作文支援に有用である事が示された。

今後の課題として、主観評価結果でも要望があった複数単語によるブーリアン検索（AND 検索と OR 検索）の実装が挙げられる。また、現在の検索システムではクエリとの完全一致での用例検索のみを行っているため、文脈の種類が限られてしまう。そこで、曖昧検索を実装することにより、文脈の種類の拡充を試みる。

謝辞

Lang-8 のデータ使用に際して、快諾してくださった株式会社 Lang-8 社長喜洋洋氏に感謝申し上げます。

参考文献

[1] 李在鎬, 淺尾仁彦, 濱野寛子, 佐野香織, 井佐原均. タグ付き日本語学習者コーパスの開発. 言語処理学会第 14 回年次大会, pp. 658–661, 2008.

[2] 曹紅荃, 八木豊, 黒田史彦, 仁科喜久子. 学習者コーパス「なたね」の構築と応用の可能性. In *CASTEL-J*, 2012.

[3] 水本智也, 小町守, 永田昌明, 松本裕治. 日本語学習者の作文自動誤り訂正のための語学学習 SNS の添削ログからの知識獲得. 人工知能学会論文誌, pp. 420–432. 2013.